Family Practice Advance Access published February 23, 2012

Family Practice 2012; **0**:1–7
doi:10.1093/fampra/cms016

Evaluating the cross-cultural validity of the Polish version of the Four-Dimensional Symptom Questionnaire (4DSQ) using differential item functioning (DIF) analysis

Sławomir Czachowski[a,*], Berend Terluin[b], Adam Izdebski[c] and Paweł Izdebski[d]

[a]Family Doctor Department, Medical College, Nicolaus Copernicus University, Torun, Poland, [b]Department of General Practice, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands, [c]Polish Academy of Sciences, Warsaw, Poland and [d]Institute of Psychology, Kazimierz Wielki University, Bydgoszcz, Poland.
*Correspondence to Slawomir Czachowski, Family Doctor Department, Medical College, Bydgoszcz, Nicolaus Copernicus University, Grabowa 10, 87–100 Torun, Poland; E-mail: s.czachowski@to.home.pl

Received 19 September 2011; Revised 5 January 2012; Accepted 2 February 2012.

Background. The original Dutch Four-Dimensional Symptom Questionnaire (4DSQ), which measures distress, depression, anxiety and somatization, has been translated into Polish with the aim of providing primary health care with a good screening instrument for the detection of the most prevalent mental health problems (anxiety, somatization, depression and distress).

Aim. To check if the Polish version is cross-culturally valid so that the scores of Polish subjects can be compared with the scores of Dutch subjects and the Dutch cut-off points can be used in Polish subjects.

Method. 4DSQ data were collected from a mixed sample of students and primary care attendees. The Polish data were compared with the 4DSQ data of a matched sample of Dutch students and primary care attendees. Two methods of differential item functioning (DIF) analysis, ordinal logistic regression and generalized Mantel–Haenszel, were used to detect items with DIF, and linear regression analysis was used to estimate the scale-level impact of DIF.

Results. Four items showing DIF were detected in the distress scale, one in the somatization scale and one in the anxiety scale. The DIF in distress caused Polish subjects with moderate scores to score circa 1 point less than their Dutch counterparts.

Conclusions. The results of the DIF analyses suggest that the Polish 4DSQ measures the same constructs as the Dutch 4DSQ and that the Dutch norms can be used for the Polish subjects, except for distress: the first cut-off point should be one point lower.

Keywords. Mental disorders, primary care, screening instrument.

Background

In primary care settings, 25–45% of patients claim mental health problems.[1] There are many tests which measure these symptoms; however, the Four-Dimensional Symptom Questionnaire (4DSQ) is one of the few instruments that aims to distinguish between distress and psychiatric disorder (anxiety and depression), which has special relevance to primary care. The 4DSQ detects the four most important symptom dimensions in primary care: distress, depression, anxiety and somatization.[2,3] The 4DSQ consists of 50 items concerning symptoms which have occurred in the past 7 days (the 4DSQ is available for download at www.emgo.nl/researchtools/4dsq.asp). The distress scale measures stress-related complaints, the depression scale measures symptoms related to depressive disorder, the anxiety scale measures symptoms connected with anxiety disorders and the somatization scale measures symptoms commensurate to bodily distress. 4DSQ test has been validated in the Netherlands and in Poland with good psychometric properties.[4] The original Dutch questionnaire was translated into Polish with back-translation procedure (S. Czachowski, A. Izdebski, B. Terluin, P. Izdebski, submitted for publication).

It is frequently assumed that merely translating the language of a questionnaire is enough for it to be used in another population. However, that may not be the case. Establishing that a translated questionnaire actually measures the same construct(s) in the same way as the original questionnaire requires the establishment of measurement equivalence. Differential item functioning (DIF) analysis provides a way to investigate whether translations of items in multi-item scales are equivalent to the original items. The probability of

endorsing an item is related to a person's position on the construct that is measured by the scale. A graphical representation of the probability of endorsing an item as a function of the scale is denoted an item characteristic curve (ICC). DIF is present when members of different (language) groups demonstrate different probabilities of endorsing an item while sharing the same position on that construct.[5] Uniform DIF indicates that the probability of endorsing an item is higher for one group over the other across the entire range of the scale, whereas non-uniform DIF is present when the probability of endorsing an item is higher for one group in one part of the scale but higher for the other group in another part of the scale. In case of uniform DIF, the ICCs for the (language) groups are completely separated, indicating that the item (translation) is more 'difficult' for one (language) group than for the other. More difficult means that a member of one group needs to occupy a higher position on the construct in order to obtain the same probability of endorsing the item as a member of the other group. In case of non-uniform DIF, the ICCs cross.

When equivalence of the Polish 4DSQ to the Dutch 4DSQ can be established, not only Polish and Dutch 4DSQ scores can be compared or pooled, but, more importantly, it can be assumed that the validity of the Dutch 4DSQ (including its cut-off points) applies to the Polish questionnaire. That would make a separate validation of the Polish 4DSQ superfluous.

Aim

Our aim was to examine if the Polish version of the 4DSQ measures the same constructs in the same way in Polish subjects as the Dutch 4DSQ measures them in Dutch subjects.

Methods

In order to compare the way the Polish and the Dutch subjects respond to the 4DSQ, two parallel samples from both cultures were created. The Polish data were collected in spring 2009 with the purpose of checking the reliability and the internal structure of the Polish version of the 4DSQ. The sample consisted of 142 students of medicine and psychology as well as 153 patients of general practice and psychiatric counselling services. Incomplete records were excluded. As for the Dutch counterparts, existing 4DSQ data were utilized. Gender-matched complete 4DSQ data were randomly selected from a larger Dutch student sample collected in 2007, and gender– and age-matched 4DSQ data were randomly selected from a larger Dutch general practice attendee's sample collected in 1993.

We decided to explore students as this group was the most accessible. It was assumed that in this population, there could occur the four dimensions of psychiatric disorders, with high intensity. Students are exposed to a lot of stress. There could be a bias, but still the datasets are comparable, as we produced the Dutch dataset in the way that fits with the Polish data.

The cross-cultural validity was assessed with the use of DIF analysis.

Given the fact that a number of different methods can be used to assess DIF and that no single method processes proven superiority over the others, Hambleton[6] recommends the use of multiple methods to detect DIF. We used a parametric and a non-parametric method. As a parametric method, we conducted ordinal logistic regression (OLR) with the use of Zumbo's SPSS syntax.[7] DIF is detected by comparing three OLR models. Model 1 models the item score as a function of the 'matching variable', for which initially the scale score was used. Model 2 models the item score as a function of the matching variable and language. Model 3 models the item score as a function of the matching variable, language and the interaction between language and the matching variable. For each model, Zumbo's syntax provides a chi-square and an R^2 value. Differences in chi-square values can be used to assess the statistical significance of DIF, while differences in R^2 values between models can be used as a measure of the effect size (ES) of DIF. If the difference between Model #1 and #3 was significant ($P < 0.001$) and the R^2 difference was >0.035, then the item was considered to show DIF.[8] The comparison of R^2 values between Models #2 and #3 allows the differentiation between uniform and non-uniform DIF. If Model #2 contributes to most of the R^2 value total difference (Model #3 minus Model #1), then the DIF is primarily uniform. In the cases of items in which the proportion of the keyed responses were too small (~10%) to compute an analysis for an ordinal score, they were recoded into binary scores and ordinary logistic regression analysis was used to determine DIF.[7] Since the presence of items with DIF in the matching variable may cause apparent DIF in another item that really does not contain DIF (this is called'pseudo-DIF') or, conversely, may conceal real DIF in another item (this is called 'hidden' DIF), the matching variables have been 'purified' throughout the analyses, i.e. items with DIF were one by one excluded from the matching variable and analyses were reiterated.

As a non-parametric method, we used the Mantel–Haenszel method.[6,8,9] The analyses were conducted with the use of J. Patrick Meyer's jMetrik software (http://itemanalysis.com/), using the scale score as the matching variable, stratification according to deciles. This strategy is to limit the number of empty cells and subsequent loss of power. The ES was estimated with the use of standardized mean difference between the

Polish and the Dutch groups, divided by overall item SD: |ES| values >0.17 were considered to indicate DIF.[10] Moreover, the statistical significance of the chi-square (d.f. = 1) value had to be <0.001 to consider an item for DIF. The matching variables were purified by omitting DIF-laden items one by one, as we did in the OLR analysis. We also produced the ICCs of each DIF-laden item for both language groups in order to see the direction of the difference between them.

To get a first impression of the impact of DIF, we simply plotted the mean scale score against an estimation of the DIF-free scale score by language. The estimated DIF-free scale score was calculated as the mean item score of the DIF-free items of the scale, multiplied by the total number of items of the scale. The DIF-free scale score was thus an approximation of the DIF-free scale score based on the information in the DIF-free items. Consistent differences in mean scale scores between the language groups in certain parts of the DIF-free scale score suggested the impact of DIF upon the scale score. Importantly, DIF in one item never affects the scale score over the whole range of the score. Rather, DIF in one item affects the scale score only in that part of the scale that corresponds with the 'difficulty' of that item. Consequently, differences in mean scale scores between the language groups suggest in which parts of the (DIF free) scale score DIF originating from the DIF-containing items might exert an impact on the scale score. For the following analyses, these parts of the DIF-free scale score where DIF might or might not impact the scale score were identified and this information was used to categorize the DIF-free scale score into a limited number of DIF-free scale score categories. Indicator variables representing these different DIF-free scale score categories were then computed. The indicator variables had a value of '1' for the DIF-free scale score category of interest and a value of '0' for all other DIF-free scale score categories. For example, if the mean scale scores plot suggested that DIF might impact the middle part of the scale, three indicator variables were computed, representing the lower, middle and higher parts of the DIF-free scale, respectively, as separate scale score categories. Importantly, the difference between the raw scale score and the estimated DIF-free scale score consisted of two components: (i) the systematic underestimation or overestimation of the scale score based on the scores of the DIF-free items of the scale and (ii) the impact of DIF in DIF-laden items on the scale score. Since only the latter component was of interest, we needed to disentangle the components. Regression analysis was an appropriate method to accomplish that objective. Therefore, linear regression analysis was used to quantify and test the scale impact of DIF in the different parts of the DIF-free scale score. For k categories of the DIF-free scale score, k linear regression analyses were performed using the differences between the raw scale score and the DIF-free scale score as dependent variable and language, DIF-free scale score categories and language with DIF-free scale score categories interaction terms as independent variables. Note that a regression model needs only k–1 DIF-free scale score categories because the last DIF-free scale score category is redundant. Note also that the regression coefficient (B) of language actually represents the difference in (raw) scale score between the language groups caused by DIF in the DIF-free scale score category that is not included, redundant, in the analysis. By repeating the analyses, each time with a different DIF-free scale score category made redundant, we obtained estimates of the mean scale impact of DIF in the various parts and categories of the DIF-free scale. Any significant scale impact of DIF was then compared with the standard error of measurement (SEM), computed as $SEM = SD\sqrt{(1-\alpha)}$ where SD represents the standard deviation of the scale score and α represents Cronbach's alpha as an estimate of the scale's reliability.

Results

The sample (N = 516) consisted of 254 Polish and 262 Dutch subjects. The proportion of students was 52% in the Polish sample and 54% in the Dutch sample. The proportion of women was 66% in both samples. The mean age was 43.5 (SD = 13.0) for Polish patients and 43.7 (SD = 12.1) for Dutch patients and 25.3 (SD = 3.9) for Polish students and 24.3 (SD = 5.6) for Dutch students. All the reliability values (Cronbach's alpha) of the Polish scales were >0.8: distress—0.881, depression—0.859, anxiety—0.823 and somatization—0.824.

DIF analyses

Distress. OLR flagged three items for DIF: #19, #20 and #39 (see Table 1). The Mantel–Haenszel method flagged four items for DIF: #17, #19, #20 and #39. Importantly, three of these items had been identified by the Zumbo method too. As for the ICC, the Polish subjects scored lower values than their Dutch counterparts (see Fig. 1).

Anxiety. With the Zumbo method, the only detected DIF showing item was Item # 27. The Mantel–Haenszel did not detect any substantial DIF. The Polish subjects scored less than their Dutch counterparts.

Somatization. Using both Zumbo and Mantel–Haenszel methods, we detected only Item # 7. Judging from the visual inspection of the ICC, the Polish subjects scored higher values than the Dutch.

Depression. No items displayed substantial DIF.

TABLE 1 *Items showing DIF*

Item number	ORL (Zumbo) method						Mantel–Haenszel method	
	R^2 step #1	R^2 step #2	R^2 step num;3	Difference R^2	Difference (type)	Items excluded from the total score	Effect size	Items excluded
7 (SOM)	0.3487	0.4210	0.4553	0.1066	Uniform	All included	0.34	All included
17 (DIS)				Not detected			0.23	#19, #20, #39
19 (DIS)	0.5693	0.5964	0.6508	0.0815	Mixed	#20 and #39	–0.28	#17, #20, #39
20 (DIS)	0.3311	0.3676	0.3813	0.0502	Uniform	#19 and #39	–0.35	#17, #19, #39
27 (ANX)	0.5495	0.5546	0.6081	0.0586	Mixed	All included	Not detected	
39 (DIS)	0.2127	0.2409	0.2485	0.0358	Mixed	#19 and #20	–0.37	#17, #19, #20

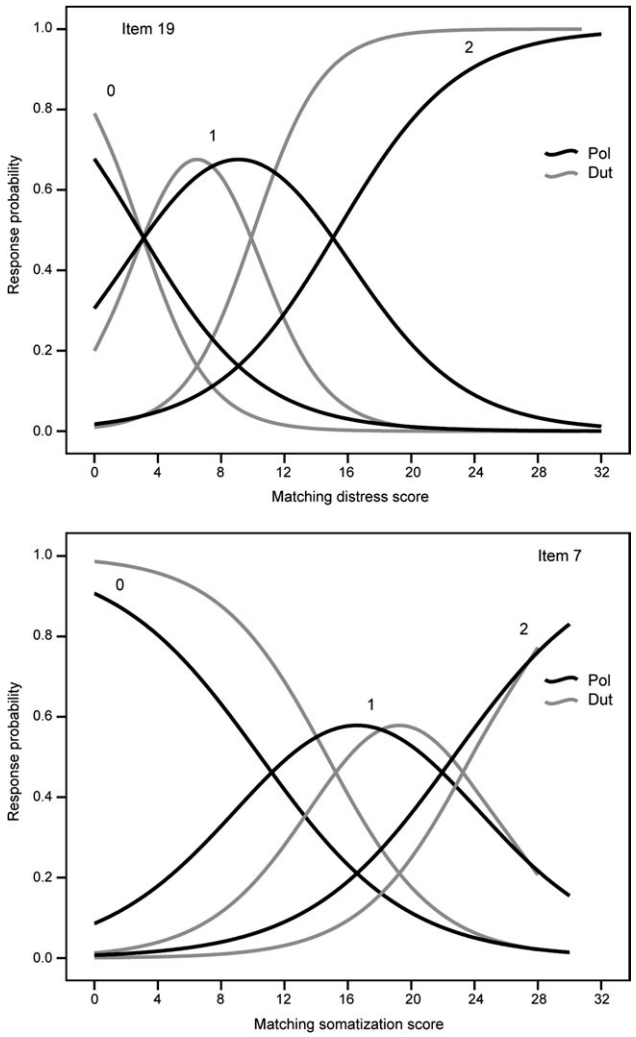

FIGURE 1 *Exemplary ICCs for two DIF-showing items: #7 (somatization) and #19 (distress). The curves show the probability of endorsing a particular item response (0 = no, 1 = sometimes, 2 = regularly or more often) as a function of the DIF-free scale score and language (Pol = Polish, Dut = Dutch)*

The scale-level impact of DIF

Distress. The Dutch cut-off points for distress are 11 and 21. After inspection of a plot of the mean distress scores as a function of the DIF-free score (Fig. 2), we divided the distress score into four categories: 0–7, 8–12, 13–19 and 20–32. In the score range around the lower cut-off point of the distress scale (8–12), DIF caused Polish people to score 1.181 points less on the distress scale ($P < 0.001$) (see Table 2). Consequently, the lower cut-off point for distress (⩾11) can be lowered to ⩾10 for Polish people to retain the same meaning as in Dutch people.

Anxiety. The cut-off points are 8 and 13. We adopted the following categories: 0–3, 4–7, 8–12 and 13–24. See Table 2 for the results. The statistically significant results were obtained for the two moderate score categories (4–12). The effect, however, was small, ∼0.4 points (lower results of the Dutch subjects for the score category 4–7, of the Polish subjects for the category 8–12), so they should not result in any changes made to the cut-off points.

Somatization. The cut-off points are 11 and 21, whereas our categories were 0–7, 8–14 and 15–32. All differences were statistically significant, yet as the effects were much lower than half a point (0.356 being the largest difference; in all score categories, the Dutch subjects scoring less than their Polish counterparts), the detected DIF cannot be considered to influence substantially the scale results.

Among the Polish sample (N = 295) for various scales, we lacked results because people did not answer all questions. These are the proportions of respondents who did not complete scales:

somatization 7.1% (21 people of 295), anxiety 4.4% (13), depression 3.1% (9) and distress 5.8% (17). In reality, the missing values are spread across all the questions. There was no difference between subjects with missing values and the complete sample.

Discussion

Even though 6 of the 50 items of the 4DSQ displayed noticeable DIF, the analyses showed that the influence of these items on total scale results is not large. Their impact is considerable only in the case of the

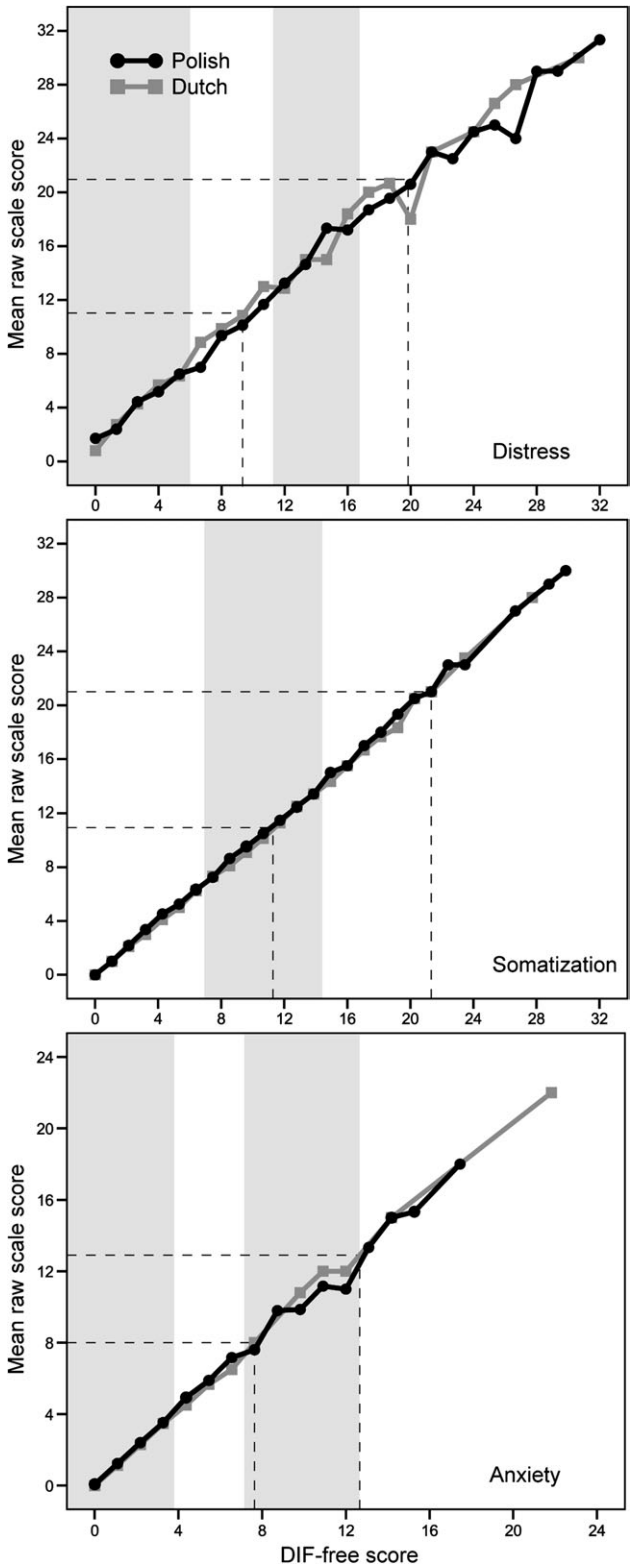

FIGURE 2 *Diagrams showing the relation between the mean raw scale score against an estimation of the DIF-free scale score by language (useful for the visual inspection of the scale impact of DIF)*

respondents scoring moderate values on the distress (Scores 7–12) scale, causing Polish people to score ~1 point less. For this reason, the first cut-off points should be lowered (11–10).

Most of the differences revealed in the data by DIF concerned the dimension of distress. Distress is a complex concept that has already been researched in primary care in the literature.[11] However, it does not

TABLE 2 *Summary table of the scale impact of DIF on the raw 4DSQ scale scores*

4DSQ scale	DIF-free score	Approximate raw score	Mean difference[a]	95% Confidence interval		*P*
				Lower bound	Upper bound	
Distress	0–5	0–7	0.028	–0.373	0.429	0.892
	6–11	8–12	–1.181	–1.728	–0.634	0.000
	12–16	13–19	0.052	–0.854	0.958	0.910
	17–32	20–32	–0.382	–1.264	0.500	0.395
Somatization	0–6	0–7	0.199	0.081	0.316	0.001
	7–13	8–14	0.216	0.076	0.357	0.003
	14–32	15–32	0.356	0.105	0.606	0.005
Anxiety	0–3	0–3	0.102	0.018	0.187	0.018
	4–6	4–7	0.418	0.086	0.750	0.014
	7–12	8–12	–0.428	–0.745	–0.110	0.008
	13–24	13–24	–0.216	–0.906	0.474	0.539

[a]Polish minus Dutch.

lend itself easily to be measured precisely as a mental state and has not been particularly in the scope of interest of primary care patients The evaluation of this phenomenon is difficult due to lack of scientific literature describing the differences in behaviour patterns in Polish and Dutch patients. Items #17 (feeling down or depressed?), #19 (worry?), #20 (disturbed sleep?) and #39 (have difficulty in getting to sleep?) express patients' emotional states. In the tradition of social discourse in Poland, they have not been used frequently.[12] Patients and doctors in Poland are not well equipped to accurately name and measure negative feelings and mental states. In the communist times, there was no possibility to publicly reveal negative emotions[13] and so any tests evaluating these disorders in medical practice were non-existent. Mental health problems in Polish primary care were published only by psychiatrists. The 4DSQ test is the first attempt in Poland to measure these characteristics. Although the patients' age indicated at a relatively young population, this group of patients were the frequent attendees at the practice and willing to participate in the research. According to our observation, it is not hazardous to posit that the content of these respondents' answers can be treated as representative for primary health care.

The differences may also stem from the tradition of disease coding in Poland, which follows the International Classification Disease (ICD-10). In Holland, the International Classification Primary Care system (ICPC) that is used takes into account a complex situation of the patient.

More specifically, in Poland, in the ICD-10 system, patients coming to the doctor are asked about axial symptoms, necessary to classify a particular disease.

In Holland, on the other hand, in the ICPC system, patients are asked about symptoms from various diagnostic axes (family background, psychological feelings, a general level of social functioning, etc.). Different ways of classifying diseases entail different styles and possibly linguistic metaphors.

In contrast to the ICPC system, ICD-10 does not appear conducive to patients' revealing their mental health problems.

Item #7 from the somatization dimension (palpitations?) has shown differences in DIF, which might be explained through the translation of the word 'palpitations' into Polish ('palpitacje'). This Polish term is more commonly used by lay people rather than professionals. It is not received by patients with such a sense of fear as other words reflecting serious cardiac problems, e.g. 'behind-sternal pain'.

In a similar vein, Item #27 from the anxiety dimension has many senses in the Polish medical language which may not possibly overlap with those of the Dutch word (angstig). In Polish, this word mainly refers to being frightened of someone rather than of something, in this case—a disease. These noticeable but minor differences in the 4DSQ using DIF call for further linguistic exploration as proposed by other studies.[14,15] It would be useful to conduct qualitative analysis of the perceptions of mental states by the Polish and the Dutch. Symptoms originating from distress or somatization have the complex mechanism of creation,[16,17] whose understanding and naming can have a sociocultural aspect.[18]

Thus, both language versions of the 4DSQ questionnaire are entirely parallel and their use in cross-cultural studies is fully justified, provided the lowered cut-off point for distress is applied.

Limitation of the study.

The limitation of this study lies in selecting a younger group of patients and a homogeneous group of students. Students, even though they are under a lot of pressure due to their educational process, are a healthier group on the whole, which does not mean though they seldom use medical services.

Conclusions

The results of the DIF analyses suggest that the Polish 4DSQ measures the same constructs as the Dutch 4DSQ. Dutch cut-off points can validly be transposed to Polish settings, except for the cut-off point for moderate distress. The outcomes explain the differences in the respondents' answers coming from culturally and linguistically diverse backgrounds and can be relevant in harmonizing the treatment of psychosomatic disorders in primary health care. In the future, they can also help in finding the most effective methods of treating these disorders. This indicates that DIF may be a useful technique in validation of new tools for and in improvement of the already implemented tests measuring the four most frequent psychiatric disorders in primary care.

Declaration

Funding: none.
Ethical approval: Ethical Committee, Kujavian-Pomeranian Doctors' Chamber in Torun. Project approval Nr 13/KB/2010.
Conflict of interest: none.

References

[1] Verhaak PFM. *Mental Disorder in the Community and in General Practice. Doctors Views and Patients Demands.* Aldershot, UK: Avebury, 1995.
[2] Terluin B, Rhenen W, Schaufeli WB, Haan M. The Four-Dimensional Symptom Questionnaire (4DSQ): measuring distress and other mental health problems in working population. *Work Stress* 2004; **18:** 187–207.
[3] Terluin B, Brouwers EPM, Marwijk van HWJ, Verhaak PFM, Horst van HE. Detecting depressive and anxiety disorders in distressed patients in primary care; comparative diagnostic accuracy of the Four-Dimensional Symptom Questionnaire (4DSQ) and the Hospital Anxiety and Depression Scale (HADS). *BMC Fam Pract* 2009; **10:** 58.
[4] Terluin B, van Marwijk HWJ, Ader HJ *et al.* The Four-Dimensional Symptom Questionnaire (4DSQ): a validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. *BMC Psychiatry* 2006; **6:** 34.
[5] Petersen MA, Groenvold M, Bjorner JB *et al*. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Qual Life Res* 2003; **12:** 373–85.
[6] Hambleton RK. Good practices for identifying differential item functioning. *Med Care* 2006; **44 (11 suppl 3):** S182–8.
[7] Zumbo BD. *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense, 1999 http://educ.ubc.ca/faculty/zumbo/DIF/ (accessed on 14 March 2010).
[8] Zumbo BD. Three generations of DIF analyses: considering where it has been, where it is now, and where it is going? *Lang Assess Q* 2007; **4:** 223–33.
[9] Budgell GR, Nambury SR, Douglas AQ. Analysis of differential item functioning in translated assessment instruments. *Appl Psychol Meas* 1995; **19:** 309–21.
[10] Laitusis CC, Maneckshana B, Monfils L, Ahlgrim-Delzell L. Differential item functioning comparisons on a performance-based alternate assessment for students with severe cognitive impairments, autism and orthopedic impairments. *J Appl test Technol* 2009; **10:** 1–33.
[11] Haller DM, Sanci LA, Sawyer SM, Patton GC. The identification of young peoples emotional distress: a study in primary care. *Br J Gen Pract* 2009; **59:** e61–70.
[12] Czachowski S, Piszczek E, Sowinska A, Hartman TC. GPs challenges in the management of patients with medically unexplained symptoms in Poland: a focus group-based study. *Fam Pract* 2011 Sep 1 [Epub ahead of print] doi:10.1093/famprj/cmr065.
[13] Czachowski S, Pawlikowska T. These reforms killed me: doctors perceptions of family medicine during the transition from communism to capitalism. *Fam Pract* 2011; **28:** 437–43.
[14] Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila Pa 1976)* 2000; **25:** 3186–91.
[15] Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol* 1993; **46:** 1417–32.
[16] Lipowski ZJ. Somatization: the concept and its clinical application. *Am J Psychiatry* 1988; **145:** 1358–68.
[17] Verkuil B, Brosschot JF, Thayer JF. A sensitive body or a sensitive mind? Associations among somatic sensitization, cognitive sensitization, health worry, and subjective health complaints. *J Psychosom Res* 2007; **63:** 673–81.
[18] Eriksen R, Holger U. Sensitization and subjective health complaints. *Scand J Psychol* 2002; **43:** 189–96.